J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr


# BART 기반 문서 요약을 통한 토픽 모델링 성능 향상☆

# Performance Improvement of Topic Modeling using BART based Document Summarization

김 은 수[1] 유 현[2] 정 경 용[1]*
Eun Su Kim Hyun Yoo Kyungyong Chung

**요 약**

정보의 증가 속에서 학문 연구의 환경은 지속적으로 변화하고 있으며, 이에 따라 대량의 문서를 효과적으로 분석하는 방법의 필요성이 대두된다. 본 연구에서는 BART(Bidirectional and Auto-Regressive Transformers) 기반의 문서 요약 모델을 사용하여 텍스트를 정제하여 핵심 내용을 추출하고, 이를 LDA(Latent Dirichlet Allocation) 알고리즘을 통한 토픽 모델링의 성능 향상 방법을 제시한다. 이는 문서 요약을 통해 LDA 토픽 모델링의 성능과 효율성을 향상시키는 접근법을 제안하고 실험을 통해 검증한다. 실험 결과, 논문 데이터를 요약하는 BART 기반 모델은 Rouge-1, Rouge-2, Rouge-L 성능 평가에서 각각 0.5819, 0.4384, 0.5038의 F1-Score를 나타내어 원문의 중요 정보를 포착하고 있음을 보인다. 또한, 요약된 문서를 사용한 토픽 모델링은 Perplexity 지표를 통한 성능 비교에서 원문을 사용한 토픽 모델링의 경우보다 약 8.08% 더 높은 성능을 보인다. 이는 토픽 모델링 과정에서 데이터 처리량의 감소와 효율성 향상에 기여한다.

☞ 주제어 : 문서 요약, BART, 토픽 모델링, LDA, Perplexity, Rouge

**ABSTRACT**

The environment of academic research is continuously changing due to the increase of information, which raises the need for an effective way to analyze and organize large amounts of documents. In this paper, we propose Performance Improvement of Topic Modeling using BART(Bidirectional and Auto-Regressive Transformers) based Document Summarization. The proposed method uses BART-based document summary model to extract the core content and improve topic modeling performance using LDA(Latent Dirichlet Allocation) algorithm. We suggest an approach to improve the performance and efficiency of LDA topic modeling through document summarization and validate it through experiments. The experimental results show that the BART-based model for summarizing article data captures the important information of the original articles with F1-Scores of 0.5819, 0.4384, and 0.5038 in Rouge-1, Rouge-2, and Rouge-L performance evaluations, respectively. In addition, topic modeling using summarized documents performs about 8.08% better than topic modeling using full text in the performance comparison using the Perplexity metric. This contributes to the reduction of data throughput and improvement of efficiency in the topic modeling process.

☞ keyword : Document Summarization, BART, Topic Modeling, LDA, Perplexity, Rouge

## 1. 서 론

시중에 유통되는 정보량의 폭발적인 증가와 그 분류의 다양화는 학문 연구의 환경을 지속적으로 바꾸고 있다. 특히, 발행되는 학술 논문의 수는 매년 빠르게 증가하며 이를 연구자들이 체계적으로 검토하고 분석하는 것은 현실적으로 어렵다. 이러한 상황 속에서 토픽 모델링은 대량의 문서에서 주제를 추출하고 조직화할 수 있다는 점에서 유용하게 사용될 수 있다. 이때 토픽 모델링은 문서 집합에서 내재된 의미를 찾고 주제를 토픽으로 군집화하는 것을 의미한다[1]. 토픽 모델링 알고리즘은 텍스트의 길이와 복잡성에 따라 성능에 제한을 받을 수 있다[2]. 학술 논문과 같이 길이가 긴 문서가 이에 해당된다.

따라서 본 연구는 토픽 모델링의 성능과 효율성 향상을 접근법을 제안한다. 그동안의 보편적인 토픽

1 Division of AI Computer Science and Engineering, Kyonggi University, Suwon, 16227, Korea.
2 Contents Convergence Software Research Institute, Kyonggi University, Suwon, 16227, Korea.
* Corresponding author (dragonhci@gmail.com)
[Received 28 November 2023, Reviewed 15 January 2024(R2 01 March 2024, Accepted 12 March 2024]
☆ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2020R1A6A1A03040583)

모델링 기법은 전체 문서를 입력으로 하여 토픽을 추출한다. 제안하는 방법은 토픽 모델링을 수행하기 전 문서에 대한 요약을 수행한다. 이후 요약문에 대한 토픽 모델링을 수행함으로써 토픽 추출의 효과와 효율성을 동시에 향상시키는 방안을 모색한다. 이를 위해 BART 문서 요약 모델을 사용하여 논문의 핵심 내용을 추출하고, 요약된 텍스트를 바탕으로 LDA 알고리즘을 통해 토픽을 추출한다. 이는 요약문을 사용하므로 전체 문서 사용 대비 적은 자원으로 토픽 추출이 가능하다. 따라서 초기 토픽 모델링 알고리즘의 학습 이후 새롭게 추가되는 논문들에 대한 요약 및 토픽 모델링 연산도 별도의 추가 학습 없이 수행할 수 있다는 장점이 있다. 또한 문서 요약 이후 토픽 모델링을 수행한다는 점은 사용자로 하여금 논문의 효율적인 검색과 이해를 도우며, 필요한 정보만을 빠르고 정확하게 찾을 수 있게 한다.

본 연구의 기여점은 다음과 같다. LDA 토픽 모델링을 수행하기 이전에 선제적인 요약을 수행함으로써 불필요한 정보를 제거하고 중요 내용을 강조할 수 있게 되어 토픽 추출 성능을 향상시킨다. 또한 요약으로 토픽 모델링에 필요한 연산량이 감소해 효율성이 향상된다.

본 논문의 구성은 다음과 같다. 2장에서 딥러닝을 활용한 문장 요약 기법과 통계적 토픽 모델링 기법에 관한 기존 연구를 기술한다. 3장에서는 BART 기반 문서 요약을 통한 토픽 모델링 성능 향상에 대해 기술한다. 이후 4장에서 실험을 수행한다. 마지막으로 5장에서는 결론을 기술한다.

## 2. 관련 연구

### 2.1 딥러닝을 활용한 문장 요약 기법

문서 요약은 문서 원본의 의미를 유지하면서도 길이를 줄이며 핵심적인 내용으로 간소화하는 과정을 의미한다. 이는 정보 검색 및 중요 내용의 강조 등을 위해 다양한 작업에서 수행된다. 문서 요약의 방법론은 크게 추출 요약과 생성 요약으로 나눌 수 있다. 추출 요약은 원본 문서에서 핵심이 되는 문장이나 단어를 추출하여 요약문을 만드는 방식이다. 반면 생성 요약은 원본 문서를 기반으로 하여 새로운 요약문을 생성하는 방식으로, 원본 문서의 의미를 보다 충실하게 반영하는 장점이 있다[3]. 추출 요약의 대표 기법으로는 TextRank가 있다. TextRank 알고리즘은 문장 사이의 유사도를 바탕으로 중요한 문장의 순위를 매긴다[4]. 한편 딥러닝의 발전으로 인해 제안된 Attention 및 Transformer 메커니즘은 생성 요약 모델의 성능 향상을 가속화한다[5]. 그러나 Transformer 아키텍처를 단독으로 사용하는 것은 기본 모델을 사용한 미세조정이 어려운 단점이 있다.

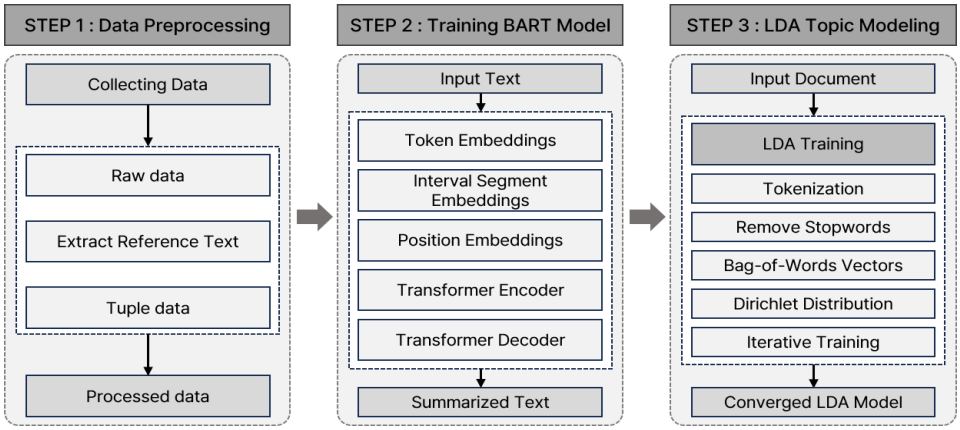
최근 연구 동향을 살펴보면, Transformer 아키텍처를 바탕으로 한 대규모 사전학습 언어 모델을 활용하는 경향이 있다. BERT(Bidirectional Encoder Representations from Transformers)는 Transformer의 인코더 구조를 사용하며, Masked Language Modeling 및 Next Sentence Prediction 방식을 통해 성능을 향상시킨다[6]. GPT는 Transformer의 디코더 구조를 사용하여 문장 생성 작업에 특화된 모델이다. 따라서 모델 구조의 개선을 통해 문장 요약 등 다양한 작업에서 좋은 성능을 보인다. 대규모 데이터를 통해 사전학습된 언어 모델은 미세 조정을 통해 필요한 훈련 데이터양과 훈련 시간을 큰 폭으로 줄이면서도 다양한 도메인과 하위 작업에 적용된다.

BART는 BERT의 인코더 구조와 GPT의 디코더 구조를 결합해 유연한 아키텍처를 구성한 언어 모델이다. BART는 문장 내 일부 토큰의 순서를 바꾸거나 무작위로 마스킹하는 등의 노이즈를 가한 뒤 원래의 문장을 복원하는 방식으로 학습된다[7]. 이러한 방식으로 인해 BART는 다양한 자연어 처리 하위 작업에서 높은 성능을 달성한다. 특히 문장 요약 작업에서 BART는 추출 요약과 생성 요약을 모두 높은 성능으로 수행할 수 있는 특징이 있다.

### 2.2 통계적인 토픽 모델링 기법

토픽 모델링(Topic Modeling)은 대량의 텍스트 데이터에서 숨겨진 의미 구조를 찾는데 사용되는 통계적 기법으로 정보 검색 및 문서 분류, 문서 요약 등 많은 자연어 처리 작업에 광범위하게 적용 가능하다. 토픽 모델링의 발전 초기에는 정보 검색 시스템의 자동 인덱싱을 활용한다[8]. 이는 키워드 매칭 기법이나 벡터 공간 모델을 사용하는 방식이 사용된다. 벡터 공간 모델은 문서와 질의를 벡터로 표현하고, 이들 간의 유사성을 계산하여 문서들 사이의 관계를 수식으로 표현한다. 하지만 이는 단어의 사용 빈도에 초점을 맞추므로 문맥적 의미를 고려하지 못하는 한계가 있다[9].

이후 문서와 질의의 관계를 확률적으로 해석하는 모델이 제안되어 문서가 질의와 적합한 정도를 확률로 제공할 수 있게 된다. 하지만 여전히 문맥적 관계나 문서 구조를 파악하는데 한계가 있다. LSA(Latent Semantic




(그림 1) BART 기반 문서 요약을 통한 토픽 모델링 성능 향상 프로세스
(Figure 1) Performance Improvement of Topic Modeling using BART based Document Summarization Process

Analysis)는 문서와 단어 간의 관계를 차원 축소를 통해 추출한다. 이는 문서 집합 내의 잠재적인 의미 구조를 찾을 수 있다는 점에서 토픽 모델링 기법의 큰 진전이라고 할 수 있다. LSA는 문서와 단어의 관계를 식별하기 위해 특이값 분해(Singular Value Decomposition, SVD)를 사용한다는 특징이 있다[8]. 이는 하나의 행렬을 세 개의 행렬 곱으로 분해하는 방법으로 LSA는 문서의 잠재적 의미를 일정 부분 추출하지만 문맥적 유사성이나 다의어를 효과적으로 처리하지 못하는 한계가 존재한다.

LSA의 개념을 확장한 pLSI(Probabilistic Latent Semantic Indexing)는 문서와 단어 사이의 관계를 확률적으로 모델링한다[9]. 각 문서가 하나 이상의 토픽으로 구성되어 있으며, 각 단어가 특정 토픽에 속할 확률을 계산하여 문서 내의 단어 분포를 나타낸다. 하지만 pLSI는 과적합의 위험성이 있으며 새로운 문서에 대한 일반화 문제가 존재한다는 한계가 있다.

## 3. BART 기반 문서 요약을 통한 토픽 모델링 성능 향상

본 연구에서는 문서의 요약을 통해 토픽 모델링의 효율성을 높이고 그 성능을 향상시키기 위해 BART 요약을 통한 텍스트 정제를 거친 뒤 LDA를 통한 토픽 모델링을 수행하는 방법을 제안한다. 그림 1은 BART 기반 문서 요약을 통한 토픽 모델링 성능 향상의 프로세스를 나타낸다. 이는 네 가지 단계로 진행된다. 첫 번째로, 논문 데이터를 수집하고 전처리하는 과정이 수행된다. 두 번째로 전처리된 데이터를 입력으로 하여 BART 문서 요약 모델을 훈련한다. 세 번째로, 요약 모델을 바탕으로 문서에 대한 토픽 모델링을 위한 요약문을 생성한다. 마지막으로 요약문을 입력으로 하여 LDA를 통한 토픽 모델링을 수행한다. 이를 통해 원문과 요약문의 토픽 모델링 성능을 비교한다.

### 3.1 데이터 수집 및 전처리

BART 기반 문서 요약을 통한 토픽 모델링 성능 향상을 위해 학술 논문 데이터를 수집한다. 해당 데이터는 AI 허브의 '논문자료 요약 데이터'를 사용한다[10]. 해당 데이터는 논문에 대한 32만 건의 생성 요약 데이터를 포함하고 있다. 이는 16만 건의 논문 전체 요약 데이터와 16만 건의 섹션별 요약 데이터를 포함하며, 한국학술지인용색인(KCI)의 Open Access 논문을 대상으로 한다. 본 연구에서는 전체 데이터 중 162,079 건의 데이터를 무작위로 선택하여 사용한다. 또한 텍스트 요약과 같이 문장들을 입력으로 하여 다른 문장들을 출력으로 하는 시퀀스-투-시퀀스(Sequence-to-Sequence) 작업을 위한 BART 미세조정 모델 학습을 위해서는 [원본 문장, 요약 문장]의 쌍으로 이루어진 튜플 형태로의 데이터 전처리가 필요하다. 수집된 원시 JSON 데이터에는 논문 ID, 저널명 등 요약 모델 학습 시에는 불필요한 데이터도 다수 포함되어 있으므로 원문 문장과 요약 문장만을 추출하여 튜플 형태로 데이터를 재구성하는 작업을 수행한다. 표 1은 요약 모델 학습을 위해 튜플 형태로 전처리된 데이터를 나타낸다.

(표 1) 요약 모델 학습을 위한 전처리된 튜플 데이터
(Table 1) Preprocessed Tuple Data for BART Summarization Model

| Original Text | Labelled Text |
|---|---|
| 본 연구의 목적은 보육교사교육원에서 수학하고 있는 예비보육교사의 ... (중략) ... 이러한 결과에 기초하여 제한점과 시사점을 논의하였다. | 이 연구의 목적은 보육교사교육원에서 수학하고 있는 예비보육교사의 수학 경험, 즉 학업성취도와 교육만족도 및 대상관계와 심리적 안녕감 간의 관련성을 알아보는 것이다. |

### 3.2 문서 요약 모델의 학습

본 연구에서는 BART 기반 모델을 사용하여 문서를 요약하는 요약 모델을 학습한다. 사용하는 기본 모델은 BART에 40GB 이상의 한국어 텍스트에 대한 추가 학습을 진행한 KoBART 모델이며, 이를 논문자료 데이터 요약에 적합하게 미세조정하는 과정을 거친다[11]. 학습 과정에서 서브워드 토큰화(Subword Tokenization)로 분할된 텍스트는 인코더에 입력되어 임베딩 및 위치 인코딩 과정을 거친다. 이후 BERT와 유사한 양방향 인코더 레이어를 거치며 어텐션(Attention) 연산이 반복 수행된다.

이후에는 GPT와 유사한 구조의 디코더를 거치며 자기 회귀적 방식의 어텐션 연산을 반복적으로 수행한다. 이 과정에서 다음 토큰을 예측하며, 각 단계의 출력 토큰은 다음 단계의 입력으로 사용될 수 있다. 이를 통해 본 연구에서는 논문 도메인에 대해 원문을 입력으로 하여 최적의 생성 요약문을 출력할 수 있도록 학습을 진행한다.

요약 모델이 생성하는 요약문은 원문 길이의 최대 0.6배를 넘지 않도록 설정한다. 해당 수치는 최적의 생성 요약문을 생성하기 위해 실험적으로 정해진 수치이다. 사용하는 데이터는 요약 모델의 학습 과정 중 성능 평가를 위해서만 사용되는 Test 데이터를 사용하며, 이 중 1,000개 문장을 임의로 선정하여 요약 작업을 수행한다.

생성된 요약문은 개행으로 나뉘어 1개의 텍스트 파일로 정제되며 토픽 모델링을 위한 LDA의 입력 데이터로 사용된다. 이를 통해 생성된 요약문과 원문 간의 LDA 토픽 모델링 결과를 비교할 수 있다. 학습된 요약 모델에 의해 생성된 요약문은 표 2와 같다.

(표 2) LDA 토픽 모델링을 위해 생성된 요약문
(Table 2) Generated Summary for LDA Topic Modeling

| Original Text | Generated Summary |
|---|---|
| 본 연구의 목적은 보육교사교육원에서 수학하고 있는 예비보육교사의 ... (중략) ... 이러한 결과에 기초하여 제한점과 시사점을 논의하였다. | 이 연구의 목적은 보육교사교육원에서 수학하고 있는 예비보육교사의 학업성취도와 교육만족도 및 대상관계와 심리적 안녕감 간의 관련성을 알아보는 것이다. |

### 3.3 토픽 모델링 성능 향상

LDA(Latent Dirichlet Allocation)는 LSA와 pLSI의 한계를 개선한 모델로, 각 문서를 디리클레(Dirichlet) 분포를 따르는 다양한 토픽으로 혼합으로 구성된다고 가정한다[12]. 디리클레 분포는 연속 확률분포의 일종으로, 여러 개의 확률 변수가 하나의 확률분포를 이룰 때 이들이 어떻게 분포될지를 모델링하는 분포이다. LDA는 토픽의 분포를 전체 문서 집합에서 학습하고, 개별 문서에 대해 해당 토픽이 혼합된 분포를 추정한다. 이를 통해 LDA는 이전보다 세밀하고 유연한 토픽 모델링 접근법을 제시한다. LDA 토픽 모델링이 수행되기 전 토큰화 및 불용어 제거 작업을 수행한다. 본 연구에서는 Mecab 형태소 분석기를 통해 명사를 추출한 뒤 두 글자 이상의 명사만을 학습에 사용한다[13]. 이후 Bag-of-Words(BoW)를 통해 문서의 단어 빈도를 나타내는 벡터를 생성한다. 이후 디리클레 분포를 사용하여 문서의 토픽 분포와 토픽의 단어 분포를 초기화하며, 문서의 토픽과 단어 분포를 추정하는 과정을 반복하여 토픽 분포를 얻는다.

## 4. 실험 및 결과

### 4.1 요약 결과 분석

본 연구에서는 학습된 문서 요약 모델의 성능 평가는 *Rouge* 지표를 통해 수행한다. 이는 *N-gram*을 통해 사람이 만든 요약문과 학습 모델이 만든 요약문의 유사도를 비교한다. 본 연구에서는 겹치는 *N-gram*의 수에 따라 *Rouge-1*과 *Rouge-2*, 제시된 요약문과의 LCS(Longest Common Sequence)를 비교하는 *Rouge-L*의 세부 지표를 사용한다. 해당 지표들은 모두 *Recall*과 *Precision*의 조화평균인 *F1-Score*를 통해 비교한다. *Rouge-N*을

계산하는 과정은 식 1과 같이 표현한다.

$$ROUGE-N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}$$

(식 1)

식 1은 라벨링된 Reference 요약문과 생성된 요약문 사이의 *N-gram Recall*을 계산하는 과정을 나타낸다. $Count_{match}(gram_n)$은 n에 따라 달라지는 *N-gram*을 측정하여 모델에서 생성된 요약에서의 *N-gram*의 개수를 세며, Reference 요약문과 생성된 요약문 모두에서 나타난 *N-gram*의 개수를 합산한다. $Count(gram_n)$은 Reference 요약문에서 *N-gram*이 나타난 횟수를 합산한다. 본 연구에서는 9713개의 데이터에 대한 성능 평가를 진행하며, 표 3은 미세 조정된 BART의 *ROUGE* 성능 결과를 나타낸다.

(표 3) 미세 조정된 BART의 ROUGE 성능 결과
(Table 3) ROUGE scores of Fine-tuned BART

| | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| Precision | 0.6201 | 0.4660 | 0.5355 |
| Recall | 0.5789 | 0.4371 | 0.5021 |
| F1-Score | 0.5819 | 0.4384 | 0.5038 |

표 3의 결과를 보면 *Rouge-1*, *Rouge-L*, *Rouge-2*의 순서로 *F1-Score*가 높게 측정된 것을 알 수 있다. 이는 단어 수준의 일치도를 측정하는 *Rouge-1*의 수치가 가장 높게 나타났음을 의미하며, 본 연구에서 사용된 학술 논문 데이터로 미세조정된 BART 요약 모델이 원본의 중요 키워드와 단어를 잘 포착함을 의미한다. 표 1의 결과에서는 *Rouge-2*의 *F1-Score*가 가장 낮게 측정되었는데, 이는 생성 요약의 불가피한 특성이라고 보여진다. 생성 요약은 원문을 바탕으로 새로운 단어들로 구성된 요약문을 생성하는 경우가 많으므로 원문과 두 단어 연속으로 일치도를 측정하는 *bi-gram* 기반의 *Rouge-2* 지표는 상대적으로 낮은 것으로 보인다. 다만 원문의 전체적인 공통 부분열을 찾아 이에 대한 일치도를 지표화하는 *Rouge-L*은 준수한 성능을 보인다.

### 4.2 토픽 모델링 결과 분석

본 연구에서는 LDA 모델의 성능을 분석하기 위해 *Perplexity(PPL)*을 사용한다. Perplexity는 모델의 일반화 성능을 평가할 수 있는 지표로, 언어 모델이 데이터를 혼동하는 정도를 의미하며 이 지표가 낮을수록 토픽 모델링 결과의 성능이 우수하다고 평가한다. *Perplexity* 지표를 계산하는 과정은 식 2와 같이 표현한다.

$$Perplexity = \exp\left(-\frac{\sum_{d \in Corpus} \sum_{w \in d} \log(p(w|d))}{\sum_{d \in Corpus} N_d}\right)$$

(식 2)

식 2에서 $p(w \mid d)$는 문서 d에 단어 $w$가 나타날 확률을 의미하며, $N_d$는 문서 d의 단어 수를 의미한다. 본 연구에서는 동일한 데이터에 대해 원문과 요약문에 대한 토픽 모델링을 수행한 뒤 *Perplexity* 수치를 비교한다. 표 4는 LDA 토픽 모델링의 *Perplexity* 성능 결과를 나타낸다.

(표 4) LDA 토픽 모델링의 Perplexity 성능 결과
(Table 4) Perplexity of LDA Topic Modeling

| | Original Text | Summarized Text |
|---|---|---|
| Perplexity | -7.9673 | -8.6111 |

표 4를 보면, 본 연구에서 학습된 요약 모델을 통해 생성된 요약문을 기반으로 한 LDA 토픽 모델링의 *Perplexity* 수치가 -8.6111로 측정된다. 낮을수록 토픽 추출 성능이 우수함을 보이는 *Perplexity*가 원문을 통한 토픽 추출 대비 약 8.08% 향상되었음을 알 수 있다.

따라서 원문을 토픽 모델링에 사용하는 것보다 선제적으로 요약한 뒤 토픽 모델링을 수행하는 것이 성능적으로 우수하며, 본 연구에서 사용된 1,000건의 데이터보다 많은 양의 데이터로 비교할 경우 토픽 모델링의 성능뿐만 아니라 토픽 모델링의 효율성 측면에서 큰 장점을 가질 것으로 기대된다.

## 5. 결 론

본 연구에서는 BART 기반 문서 요약을 통한 토픽 모델링 성능 향상 방안을 제안하고 실험을 통한 검증을 수행하였다. 제안된 방법은 기존의 문서 전체를 거친 토픽 모델링 접근법과는 달리, 선제적인 문서 요약을 통해 원문의 핵심 내용을 축약하여 토픽 모델링의 입력으로 사용한다. 이를 위해 BART 기반의 문서 요약 모델을 훈련하고, 요약된 문서를 LDA 알고리즘을 통해 토픽 모델링하는 과정을 거쳤다. 실험 결과, 미세조정된

BART를 통해 생성된 요약문 기반의 LDA 토픽 모델링이 원문 기반의 토픽 모델링에 비해 우수한 *Perplexity* 성능을 보였다. 이는 요약 과정이 문서 내 중요 내용의 강조 및 불필요한 정보의 제거를 도와, 토픽 모델링이 더 명확하고 정확한 주제 분포를 도출할 수 있도록 한 것으로 보인다. 또한, 요약문을 사용함으로써 처리해야 할 데이터의 양이 감소하여 토픽 모델링의 효율성도 향상되었다. 따라서 논문 데이터를 활용하는 분야뿐만 아니라 토픽 모델링을 활용하는 다양한 분야에서 의미 있는 시사점을 제공한다. 특히, 대규모 문서 데이터를 다루는 연구에서 본 방법은 연구자들이 빠르고 정확하게 원하는 정보를 파악하는 데 도움이 될 것으로 기대된다.

## 참고문헌(Reference)


[1] P. Kherwa and P. Bansal, "Topic modeling: a Comprehensive Review," EAI Endorsed transactions on scalable information systems, 2019. http://dx.doi.org/10.4108/eai.13-7-2018.159623

[2] J. Qiang, Z. Qian, Y. Li, Y. Yuan and X. Wu, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," IEEE Transactions on Knowledge and Data Engineering, Vol. 34, No. 3, pp. 1427-1445, 2022. http://dx.doi.org/10.1109/TKDE.2020.2992485

[3] WS. El-Kassas, CR. Salama, AA, Rafea and HK. Mohamed, "Automatic Text Summarization: A Comprehensive Survey," Expert Systems with Applications, Vol. 165, 113679, 2021. https://doi.org/10.1016/j.eswa.2020.113679

[4] H. Yoo, R. C. Park, and K. Chung, "IoT-Based Health Big-Data Process Technologies: A Survey," KSII Transactions on Internet and Information Systems, Vol. 15, No. 3, pp. 974-992, 2021. https://doi.org/10.3837/tiis.2021.03.009

[5] B. Jeon, K Chung, "CutPaste-Based Anomaly Detection Model using Multi Scale Feature Extraction in Time Series Streaming Data," KSII Transactions on Internet and Information Systems, Vol. 16, No. 8, 2022. http://dx.doi.org/10.3837/tiis.2022.08.018

[6] J. Delvin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of naacL-HLT, Vol. 1, pp. 4171–4186, 2019. http://dx.doi.org/10.18653/v1/N19-1423

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and Luke Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, 2020. http://dx.doi.org/10.18653/v1/2020.acl-main.703

[8] H. Yoo, K. Chung, "Deep Learning-based Evolutionary Recommendation Model for heterogeneous Big Data Integration," KSII Transactions on Internet and Information Systems, Vol. 14, No. 9, pp. 3730-3744, 2020. https://doi.org/10.3837/tiis.2020.09.009

[9] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," Expert Systems with Applications, Vol. 42, pp. 5645-5657, 2015. https://doi.org/10.1016/j.eswa.2015.02.055

[10] AI Hub, [Online] : https://aihub.or.kr/, 2023.

[11] KoBART, [Online] : https://github.com/SKT-AI/KoBART, 2023.

[12] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, 78, 15169-15211, 2019. https://doi.org/10.1007/s11042-018-6894-4

[13] K. Park, J. Lee, S. Jang, and D. Jung, "An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks," Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 133-142, 2020. https://doi.org/10.48550/arXiv.2010.02534

## ◐ 저 자 소 개 ◑

**김 은 수(Eun su Kim)**
2019년~현재 경기대학교 AI컴퓨터공학부 학사과정 재학
2023년~현재 경기대학교 AI컴퓨터공학부 데이터마이닝 연구실 학생연구원
관심분야 : 인공지능, 데이터마이닝, 자연어 처리, 빅데이터, 머신러닝
E-mail : kimes0228@gmail.com

**유 현(Hyun Yoo)**
1999년 상지대학교 전산학과 (이학사)
2011년 상지대학교 컴퓨터교육학과 (교육학석사)
2019년 상지대학교 컴퓨터정보공학과 (공학박사)
2020년~현재 경기대학교 콘텐츠융합소프트웨어 연구소 연구교수
관심분야 : 딥러닝, 인공지능, 빅데이터 마이닝
E-mail : rhpa0916@gmail.com

**정 경 용(Kyungyong Chung)**
2000년 인하대학교 전자계산공학과(공학사)
2002년 인하대학교 전자계산공학과(공학석사)
2005년 인하대학교 컴퓨터정보공학부(공학박사)
2006년~2017년 상지대학교 컴퓨터정보공학부 교수
2017년~현재 경기대학교 AI컴퓨터공학부 교수
관심분야 : 데이터마이닝, 헬스케어, 빅데이터, 지능시스템, 인공지능, HCI, 정보검색, 추천 시스템
E-mail : dragonhci@gmail.com